DOCUMENT RESUME

ED 341 735 TM 017 931

AUTHOR Muthen, Bengt O.
TITLE Multilevel Factor Analysis of Class and Student Achievement Components.
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY National Science Foundation, Washington, D.C.; Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-332
PUB DATE Oct 90
CONTRACT OERI-G-86-003; SES-8821668
NOTE 49p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Analysis of Variance; *Factor Analysis; Grade 8; International Studies; Junior High Schools; *Junior High School Students; Mathematical Models; *Mathematics Achievement; Mathematics Tests; Pretests Posttests; *Sample Size
IDENTIFIERS Between Group Differences; Decomposition Analysis (Statistics); *Multilevel Analysis; *Second International Mathematics Study; Variance (Statistical); Within Group Differences

ABSTRACT

Issues related to between-class and within-class decomposition of achievement variance and the change of this decomposition over the course of the eighth-grade were examined using the Second International Mathematics Study (SIMS, 1985), a study in which there was a nested or hierarchical data structure of students within classes, within schools, in school districts. The usefulness of multilevel factor analysis (MFA) was explored with a subset of data containing 3,724 eighth graders in the United States from about 200 classes in about 100 schools. The core test for analysis consisted of 39 items in arithmetic, algebra, geometry, and measurement given as a pretest and posttest in fall 1982 and spring 1983. It was found that the strong elements of tracking in eighth-grade mathematics classes make for between-class variation that is about as large as the within-class student variation. Within-class variability increases much more substantially than between-class variation over the course of the eighth-grade. MFA is considered to give better results than analysis of variance or conventional factor analysis. However, MFA calls for data with a sizable number of groups, preferably at least 100. Extensions of the MFA model are briefly discussed. Seven tables of study data, 2 figures, and a list of 27 references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED341735

C·R·E·S·S·T

Center for Research
on Evaluation, Standards,
and Student Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

MULTILEVEL FACTOR ANALYSIS OF CLASS AND STUDENT ACHIEVEMENT COMPONENTS

CSE Technical Report 332

Bengt O. Muthen

UCLA Center for the Study of Evaluation

TM017931

BEST COPY AVAILABLE

MULTILEVEL FACTOR ANALYSIS OF CLASS AND STUDENT ACHIEVEMENT COMPONENTS

CSE Technical Report 332

Bengt O. Muthen

UCLA Center for the Study of Evaluation

October 1990

This research was supported by grant OERI-G-86-003 from the Office of Educational Research and Improvement, Department of Education and by the National Science Foundation grant SES-8821668. I would like to thank Ginger Nelson, Jin-Wen Yang Hsu, Kathleen Wisnicki, and Tammy Tam for valuable research assistance and my Fall 1990 research seminar group of stimulating discussions.

1. Introduction

Educational research often depends on multivariate techniques like confirmatory factor analysis and other covariance structure techniques (see, e.g. Joreskog & Sorbom, 1979; Bollen, 1989) to study underlying dimensions of systematic variation in student data and to assess reliability and invariance of measurement instruments (see, e.g. Bohmstedt, 1983). Standard analysis methods make the simplifying assumption that the student data have been obtained as a simple random sample from a given population. This involves an assumption of independently and identically distributed observations. Much educational data collection, however, is obtained through a complex, multi–stage sample design involving clustered observations where this assumption is unrealistic. Typical examples are large–scale surveys like the often–used National Longitudinal Study (NLS) of the high school graduates of 1972 and the newly launched National Longitudinal Study of 1988 (NELS:88) with stratified sampling of schools and random sampling of students within schools. This paper considers another large–scale survey, the Second International Mathematics Study (SIMS), with a similar nested or hierarchical data structure of students observed within classes, obtained within schools within school districts. Standard analysis methods are adversely affected by such deviation from simple random sampling (see, e.g. Skinner, Holt, Smith, 1989). For achievement tests in schools the violation of the assumption of independent observations may be particularly important since students of the same class are likely to produce sizeable intraclass correlations due to strong common sources of variation. In the last few years suitable techniques have become popularized for univariate response models including multiple regression using random coefficient or multilevel regression models (see, e.g. Bock, 1989; Raudenbush & Bryk, 1988). Extensions of techniques for multivariate response models are, however, just emerging. Some new developments of this kind will be demonstrated in this paper.

From the above sampling perspective design features can be viewed as complicating the statistical analysis. The complex sample design of educational studies, however, also offers an opportunity for more informative modeling of substantive phenomena. There is often an explicit interest in relating the variation in the data to the multiple stages of the sampling, such as school, class, and student. Data is often gathered on such multiple levels or units of

observation and there is an interest in studying the interaction among these levels. For example in NELS:88 there is not only an interest in student–level data but also data obtained for these students' teachers, school principal, and parents. SIMS has information on eighth and twelfth grade mathematics achievement, where effects of differing amounts of "tracking" in different educational systems in different countries and provinces can be studied. In SIMS it is therefore of interest to separate within–class and between–class variation of student achievement, to relate between–class achievement variation to class–level information on teacher and teaching characteristics, and to contrast different educational systems (see, e.g. Burstein, 1990).

While the statistical concerns about multivariate modeling in complex samples and the emerging solutions are rather new, the substantive concerns about "multilevel" modeling are relatively old including issues related to the proper unit of analysis, aggregation effects, and contextual effects. For overviews in educational and sociological contexts, see Cronbach (1976) and Burstein (1980). The Cronbach reference is particularly relevant to this paper since he discusses issues of factor analysis on multiple levels. Cronbach reanalyzes Bond–Dykstra data from the Cooperative Reading study using separate factoring of within–class and between–class covariance matrices for ability measures. Harnqvist (1978) uses a similar approach to factor analyze mental ability scores of students observed within classrooms. These analyses point to different structures on the two levels and when using the usual overall covariance matrix. However, these ideas do not appear to have had a large impact on factor analysis practice when it comes to hierarchical data such as students within classes. One reason is perhaps that the statistical methodology and software development has lagged behind (see, however, Schmidt, 1969).

Relevant statistical methodology is now emerging for efficient multivariate analyses of the kind that Cronbach and others envisioned (for an overview, see Muthen, 1989). The aim of this paper is to address some substantive analysis questions in the SIMS data and let these analyses indicate the considerable potential of this new methodology. Mathematics achievement for U.S. eighth graders will be studied. These students are to some extent selected into different types of eighth grade math classes based on previous performance. Typically, arithmetic content is well covered, but there are major differences in how much algebra and geometry is taught. In this way, the classes can be characterized in

broad categories like remedial, typical, algebra, and enriched, but differences in emphases across classes remain even within these categories. This paper studies issues related to between– and within–class decomposition of achievement variance and the change of this decomposition over the course of the eighth grade.

2. The data

In the Second International Mathematics Study (Crosswhite, Dossey, Swafford, McKnight & Cooney, 1985) a national probability sample of school districts was selected proportional to size; a probability sample of schools was selected proportional to size within school district; and two classes were randomly drawn within each school. We will consider a subset of the U.S eighth grade data concerning 3,724 students who took the core test at both the pretest in Fall of 1982 and posttest in Spring of 1983. These students are observed in about 200 classes from about 100 schools. The class sizes vary from 2 to 38 with a typical value around 20.

The core test consisted of 40 items in the areas of arithmetic, algebra, geometry, and measurement. For more detail these topics will be further broken down resulting in eight subscores to be analyzed, where each subscore is obtained as the sum of right–wrong scored items taken from the 40 core items. One item in the core had a very low item–test correlation and was excluded, so 39 items were used in total.

The subscore RPP consists of eight ratio, proportion, and percent (RPP) items. FRACT consists of eight common and decimal fraction items. EQEXP consists of six algebra items involving equalities and expression. INTNUM consists of two items involving integer number algebra manipulations. STESTI consists of five items dealing with measurement items involving standard units and estimation. AREAVOL consists of two measurement items dealing with area and volume determination. COORVIS consists of three geometry items involving coordinates and spatial vizualization. PFIGURE consists of five geometry items involving properties of plane figures.

Although the subscores consist of relatively few items and may be rather unreliable, it is of interest to be able to separately study these variables since to

some extent they correspond to different emphases in eighth grade mathematics curricula. Later on we will compare the analysis of these eight items with the analysis of four aggregates of these variables, resulting in Arithmetic, Algebra, Measurement and Geometry scores.

Teacher–reported opportunity to learn (OTL) information was also recorded for these items. For each item the value 0 or 1 was recorded, where 1 was given if the mathematics needed to solve the item had been taught during eighth grade or in prior years. The achievement subscore averages and corresponding OTL averages for these eight variables are given in Table 1 for both the pretest and posttest occasion. We see that OTL varies considerably over subscores. For the arithmetic topics of RPP and FRACT the OTL variable obtains close to a maximum score of 8, while for the algebra topic of EQEXP and the geometry topic of PFIGURE only about 2/3 of the maximum possible OTL value is observed. For EQEXP and PFIGURE the OTL also has a relatively large standard deviation, corresponding to the tracking effect of different classes putting different emphasis on algebra and geometry training.

Insert Table 1

Of particular interest in this paper is the variance decomposition of the subscores with respect to within–class student variation and between–class variation, and the change of this decomposition from pretest to posttest. For related SIMS analyses, see Schmidt, Wolfe, & Kifer (1990). For other test score analyses of this type, see e.g. Wiley and Bock (1967) and Rakow, Airasian, and Madaus (1978). In the SIMS such variance decomposition relates to effects of tracking and differential curricula in eighth grade math, where one may hypothesize that such selection effects tend to increase between–group variation at the expense of within–group variation. We will focus on the relative amount of between– and within–class variation. However, since the data hierarchy involves school, class, and student we will first consider this three–level decomposition using a standard random effects nested analysis of variance model (see e.g. Winer, 1971)

(1) $y_{ghi} = \mu + \alpha_g + \beta_{gh} + \gamma_{ghi}$,

for individual i observed within the h^{th} class within the g^{th} school. Here, μ is the overall mean and α, β, γ are independent random normal variables with zero means and variances to be estimated. The variance estimates are obtained for each of the eight subscores at both pretest and posttest using BMDP3V's maximum likelihood estimator for unbalanced, nested data.

Table 2 gives the variance decomposition in terms of percentages. All variance estimates are significantly different from zero except the pretest school variances, for which only STESTI is significant. The estimates show that the within–class, student–level percentage of the variance clearly dominates the subscore variability. The within variation is about 60 – 80% while the between variation is divided into about 20 – 30% for classes and about 3 – 13 % for schools. Both within variation and between variation increases over time.

Table 2

The two right–most columns of Table 2 give the difference of the posttest and pretest value relative to the pretest value for between (school and class) and within. This shows that the between components increase much more strongly than the within components. The between variation increase is particularly strong for the algebra content of EQEXP and the geometry content of PFIGURE. This is in line with the OTL variation across classes discussed in conjunction with Table 1. In terms of percentage of total variation, however, the within variation decreases only an average of about 5 – 10% over time. This suggests that the heterogeneity within classes remains very large. The influence of tracking on between–class variation in math achievement makes for a strong increase in between variation over eight grade. Within variation still clearly dominates although it does not increase much over time.

We should, however, note that the within variation includes individual–level measurement error variance which would inflate the contribution of with n

variation. Table 2 shows that the student percentage of the variance is the lowest for the two subscores created from the largest number of items, RPP and FRACT. It may be that this is an artifact of the possibly higher reliability of these two subscores. Also, the relative size of the measurement error is presumably larger at pretest than at posttest since less learning has taken place at pretest. This would confound comparisons over time of relative variance contributions. The issue of how measurement error can be taken into account will be considered next.

3. A multilevel factor model

Each of the eight achievement subscores are created by the summing of rather few dichotomous items. In this way they are all likely to contain a sizeable amount of measurement error. At the same time the eight achievement subscores pertain to various aspects of central eight grade math topics. Taken together this points to the use of a multivariate measurement model in terms of a factor–analytic, multiple–indicator model for the eight subscores.

3.1 Variance decomposition

Consider a variance component decomposition of the eight–dimensional observation vector $\mathbf{y}$ for individual i in group (class) g

$$(2)\quad \mathbf{y}_{gi} = \mathbf{y}_{Bg} + \mathbf{y}_{Wgi},$$

where the between component $\mathbf{y}_{Bg}$ and the within component $\mathbf{y}_{Wgi}$ are independent as in conventional random effects analysis of variance. The between component contains class and school contributions to the individual's score while the within component represents the contribution of the student. For simplicity we will not separate the school and class components here. In this way the variance of $\mathbf{y}_{gi}$ can be decomposed into a between and a within part,

$$(3)\quad \mathbf{\Sigma} = \mathbf{\Sigma}_B + \mathbf{\Sigma}_W.$$

In the present application we will assume that a one–factor model holds for both the between and the within components of (2) and (3). For a given variable j in the vector $\mathbf{y}_{gi}$ we may therefore decompose into four independently varying parts,

(4) $y_{gij} = \nu + \lambda_{Bj}\,\eta_{Bg} + \varepsilon_{Bgj} + \lambda_{Wj}\,\eta_{Wgi} + \varepsilon_{Wgij}$,

where ν is an intercept parameter and the λ's are loading parameters.

The within part of (4) can be interpreted in line with conventional factor analysis in that the within factor η_{Wgi} and the within residual ε_{Wgi} refer to individual–level variation. In the multilevel model this is within–group variation. The single factor accounts for all covariation among the individual–level achievement scores, representing a general math achievement trait for these eighth graders. While other research on these data suggest several minor factors on the item level (see, e.g. Muthen, 1988), these factors largely vanish in the aggregated scores. The residuals are viewed as measurement errors, that is variable–specific individual variation not accounted for by the factor. These errors are independent of the factor and are independent of each other.

The between part of (4) departs from conventional analysis in that it addresses across–group variation rather than across–individual variation. Here the factor η_B is interpreted in terms of selection effects due to tracking and differences in curricula. The single factor represents a single dimension on which selection is made with differing λ_B coefficients giving different weigths to different topics. One may for example hypothesize that entering eighth grade selection is dominated by previous arithmetic performance so that the between loadings are relatively higher for RPP and FRACT. During eight grade curricula vary greatly in terms of both algebra and geometry topics. This means that at posttest the variance of the between residual ε_B may increase for these topics and the unidimensionality of the between factor model may be called into question.

Consider now the variance components of (4) for observed variable j.

(5) $\sigma^2_{ygij} = \lambda^2_{Bj}\,\sigma^2_{\eta B} + \sigma^2_{\varepsilon Bj} + \lambda^2_{Wj}\,\sigma^2_{\eta W} + \sigma^2_{\varepsilon Wj}$

$= \quad BF \quad + \; BE \; + \quad WF \quad + \; WE\,,$

say, where B and W stand for between and within, while F and E stand for factor and error.

3.2 Reliability

The conventional factor analysis definition of reliability uses the R^2–like ratio of variance due to the factor divided by total variance in y. Since this refers to individual–level reliability the analogous reliability definition for y_j in multilevel data in our model would be

(6) Within reliability $(y_j) = WF \;/\; (WF + WE)$,

while the reliability in the across–group variation is

(7) Between reliability $(y_j) = BF \;/\; (BF + BE)$.

3.3 Intraclass correlation

Analogous to random effects analysis of variance (Winer, 1971) we note that the model implies the following correlation between two individuals i and i' in group g for variable y_j,

(8) $\text{Corr}\,(y_{gij}, y_{gi'j}) = \text{Cov}\,(y_{gij}, y_{gi'j}) \;/\; \sigma^2_{ygij}$

$= (BF + BE) \;/\; (BF + BE + WF + WE).$

In this model this intraclass correlation (Fisher, 1958; Haggard, 1958; Koch,

1983) is also the proportion of between variance in y_j. The larger it is the further we deviate from the conventional assumption of all observations being independent. If BF and BE are both zero for all variables there is no need for a multilevel analysis and the independence assumption of a conventional analysis are fulfilled. As we saw in Table 2 the between variance components appear sizeable in this data set. The proportions of between variance are obtained from Table 2 by subtracting the student percentage of variance from 100. This gives a proportion of about 0.2 – 0.4. Again, this proportion is influenced by measurement error as is clear from (8).

Given the decomposition in (5) into factor and error variance we may consider an error–free version of the variance ratio in (8), namely the factor variance ratio or "true intraclass correlation coefficient" for each variable,

(9) BF / (BF + WF).

In Table 2 we also considered the percentage of variance increase from pretest to posttest for between and within. This increase can also be presented in an "error–free" form using pre– and posttest values for BF and WF, respectively.

4. Multivariate multilevel estimation

The above modeling leads to a covariance structure model for two–level data which uses a conventional factor analysis covariance structure on both the between and within level,

(10) $\Sigma_B = \Lambda_B \Psi_B \Lambda_B' + \Theta_B$,

(11) $\Sigma_W = \Lambda_W \Psi_W \Lambda_W' + \Theta_W$,

where the Λ matrices contain the loadings λ_{Bj} and λ_{Wj} and have one column in this application, the Ψ matrices represent the factor variances $\sigma^2_{\eta B}$ and $\sigma^2_{\eta W}$, and the Θ matrices represent the covariance matrices for the residuals ε_{Bj} and ε_{Wj}. Muthen and Satorra (1989) and Muthen (1989, 1990) consider variations of

multilevel models leading to these and related covariance structures. These papers also show the relationship of these models to random parameter models that have become popular in educational applications of regression analysis (see, e.g. Raudenbush & Bryk, 1988). Essentially, the above modeling may be viewed as a random factor means, random measurement intercepts model. Random slopes are not involved. The modeling can be extended to 3–level data for school, class, and student but that will not be considered here.

Maximum likelihood estimation of multilevel covariance structure models was studied already by Schmidt (1969), see also Schmidt and Wisenbaker (1986), but these techniques do not appear to have gotten into practical use. More recent contributions are Goldstein and McDonald (1988), McDonald and Goldstein (1989), Longford and Muthen (1990), and Muthen (1989, 1990). The technical details of this estimation will not be presented here. Briefly stated, in the two–level case with G groups the likelihood is considered for G multivariate normal observation vectors, where each vector contains all variables for all individuals in the group. There are N_g individuals in group g, where $N = \Sigma N_g$ is the total sample size. Unlike conventional analysis independence of observations is not assumed for all N observations but only over the G groups, while the intraclass correlation is modelled via $\mathbf{\Sigma}_B$. The covariance matrices of $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$ contain the parameters of interest. In this paper we will assume that we study the common case of no mean structure. As opposed to conventional covariance structure analysis we do not only use the regular p x p sample covariance matrix, but more sample information is available. In the balanced case the ML procedure leads to the use of the customary pooled–within and between sample covariance matrices. A large–sample chi–square variable is obtained to test restrictions imposed by the model on $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$. With p variables and r parameters the number of degrees of freedom is $p(p+1) - r$. We note that a conventional covariance structure model has $p(p+1)/2 - r$ degrees of freedom since this analysis restricts the matrix $\mathbf{\Sigma}_B$ to be zero (in this case r is reduced by the number of parameters for the between part). For more details and relations to conventional structural equation modeling, see Muthen (1989, 1990).

While in principal special formulas and software could be developed for multilevel factor analysis (MFA) maximum–likelihood (ML) estimation, Muthen (1989, 1990) showed that multiple–group structural equation software can be

modified for MFA ML analysis. In line with this idea, Muthen (1990) proposed a simpler ML–based MFA estimator which can be used with already existing multiple–group structural equation software such as LISREL, LISCOMP, and EQS. In the balanced case this estimator is the MFA ML estimator. In the unbalanced case the estimator is still consistent and the chi–square test of model fit and standard errors of estimates can be used as rough approximations of the ML values. We will use both procedures in our analyses for comparison purposes. The true ML procedure will be referred to as FIML (full information ML) and the simpler estimator as MUML (Muthen's ML–based estimator). Muthen (1990) found no important differences between the results of FIML and MUML.

The MUML estimator demonstrates the basic features of MFA. Consider the three customary sample covariance matrices $\mathbf{S}_T$, $\mathbf{S}_{PW}$, $\mathbf{S}_B$,

$$(12)\quad S_T = (N - 1)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (y_{gi} - \bar{y}) (y_{gi} - \bar{y})'$$

$$(13)\quad S_{PW} = (N - G)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (y_{gi} - \bar{y}_g) (y_{gi} - \bar{y}_g)'$$

$$(14)\quad S_B = (G - 1)^{-1} \sum_{g=1}^{G} N_g (\bar{y}_g - \bar{y}) (\bar{y}_g - \bar{y})'$$

The matrix $\mathbf{S}_T$ is used in conventional covariance structure analysis. In the multilevel case it is a consistent estimator of the total covariance matrix $\mathbf{\Sigma}_B + \mathbf{\Sigma}_W$. The pooled–within matrix $\mathbf{S}_{PW}$ is a consistent and unbiased estimator of $\mathbf{\Sigma}_W$, while the between matrix $\mathbf{S}_B$ is a consistent and unbiased estimator of

$$(15)\quad \mathbf{\Sigma}_W + c\, \mathbf{\Sigma}_B,$$

where c reflects the group size,

$$(16)\ c = \left[N^2 - \sum_{g=1}^{G} N_g^2 \right] [N (G - 1)]^{-1}$$

(Muthen, 1990). For balanced data c is the common group size. For unbalanced data c is often close to the mean of the group sizes. Note that the between matrix $\mathbf{S}_B$ is the covariance matrix of group means weighted by the group size.

The MUML MFA estimator (Muthen, 1990) considers the ML–like fitting function

$$(17)\quad \begin{aligned} & G \{ \ln|\Sigma_W + c\,\Sigma_B| + \text{trace}\,[\Sigma_W + c\,\Sigma_B]^{-1} S_B - \ln|S_B| - p\} + \\ & (N - G) \{ \ln|\Sigma_W| + \text{trace}\,[\Sigma_W^{-1} S_{PW}] - \ln|S_{PW}| - p\} \end{aligned}$$

This fitting function is analogous to a two–group covariance structure ML estimator where $\mathbf{S}_B$ and $\mathbf{S}_{PW}$ are used to fit their corresponding population quantities. The first group has G "observations" while the second has N – G observations. In the balanced case this is the MFA FIML estimator in the common case of an unrestricted mean vector. The divisor is then G instead of G – 1 for $\mathbf{S}_B$. The estimation of MFA parameters via (17) can be performed by the ML fitting function in conventional multiple–group structural equation software. This fitting function automatically gives the pseudo chi–square test of model fit of H_0 against unrestricted Σ_B and Σ_W matrices as is desired. The $\mathbf{S}_B$ and $\mathbf{S}_{PW}$ matrices can be obtained via standard statistical packages. The author has written a program, available to anyone who wants it, which computes these two matrices, the c value, and the intraclass correlations for two–level data. This means that the MUML estimator is easily accessible today, while this is not true for FIML. The FIML estimator uses a fitting function similar to (17), but involves terms for each distinct group size, including information on the mean vectors (Muthen, 1990). Even when FIML can be done it will be computationally

heavier than MUML when the number of distinct group sizes increases.

Muthen (1990) showed that the input specification for the structural equation model software needed for MUML using (17) can be conveniently indicated via conventional path diagrams. Using a one–factor model for both between and within leads to the model diagram of Figure 1. This diagram follows the notation

---

Insert Figure 1

---

of (4). Below the row of squares are variables on the within level, ε_W and η_W. This part of the diagram corresponds to a conventional one–factor model. Above the row of squares is a row of circles corresponding to the between components, y_B. In this way, the observed variables y in the squares are functions of within and between components. The between components follow a one–factor model with residuals ε_B and factor η_B.

The path diagram corresponds directly to the first group in the two–group setup indicated by (17). The first group involves the covariance matrix structure $\Sigma_W + c\,\Sigma_B$. This deviates from the total covariance matrix $\Sigma_W + \Sigma_B$ by the scalar multiplier c for the between part. This means that the between components of the variables have to be scaled by $\sqrt{c}$ which is accomplished by letting the paths (loadings) from the y_B's to the y 's have coefficients $\sqrt{c}$. The second group in (17) corresponds to the within variation. The covariance structure of Σ_W is captured by using the same model structure as for the first group, following Figure 1, but fixing all between coefficients and variance–covariance parameters to zero. Since Σ_W also appears in the covariance structure of the first group equality restrictions across groups need to be applied for the within parameters.

As pointed out in Muthen (1989), MFA is a complex analysis which needs to follow a sound analysis strategy. The actual MFA should in a typical case be preceeded by four important analysis steps, conventional analysis of S_T, estimation of size of between variation, conventional analysis of S_{PW}, and

conventional analysis of $\mathbf{S}_B$.

(1) Conventional factor analysis of $\mathbf{S}_T$. This analysis is useful to try out model ideas. The analysis is incorrect when the data is multilevel due to the correlated observations. The model test of fit is usually inflated, particularly for data with large intraclass correlations, large class sizes, and highly correlated variables. However, the test of fit may still be of practical usefulness to give a rough sense of fit.

(2) Estimation of size of between variation. It is wise to first check if a multilevel analysis is warranted. This can be carried out in an MFA by testing $\mathbf{\Sigma}_B = 0$, but a simpler way to get a rough indication of the amount of between variation is to compute the estimated intraclass correlations for each variable obtained as the ML estimate of

(18) $\sigma^2_B / (\sigma^2_B + \sigma^2_W)$

where σ^2_W is estimated as s^2_{PW} and σ^2_B is estimated as

(19) $c^{-1} (s^2_B - s^2_{PW})$

(see also Winer, 1971; Muthen, 1990). If all intraclass correlations are close to zero it may not be worthwhile to go further.

(3) Conventional factor analysis of $\mathbf{S}_{PW}$. If the multilevel model is correct, a conventional analysis of $\mathbf{S}_{PW}$ is the same as an MFA with an unrestricted $\mathbf{\Sigma}_B$ matrix. This analysis estimates individual–level parameters only. Experience has shown that the analysis gives estimates that are close to the within parameters of an MFA. The conventional analysis would use a sample size of N – G and either the normal theory GLS or ML estimator. The $\mathbf{S}_{PW}$ analysis is expected to give a better model fit than the $\mathbf{S}_T$ analysis.

(4) Conventional factor analysis of $\mathbf{S}_B$. Little may be known about the factor

structure of Σ_B since it does not concern the customary individual–level data but instead across–group variation. The between components have different meaning than the within components. As the Cronbach (1976) and Harnqvist (1978) analyses showed, the same structure as on the within level cannot be counted on. To explore the between structure it is tempting to use S_B. Note, however, that S_B is not an unbiased or consistent estimator of Σ_B as is indicated in both (15) and (19). The Σ_B estimator is also a function of S_{PW}. Equation (19) generalizes directly to the multivariate case to show this estimator. In other words, any simple structure expected to hold for Σ_B does not necessarily hold for S_B but it should hold within sampling error for the ML estimate of Σ_B. Unfortunately, the ML estimator of Σ_B is frequently not positive definite and may not even have positive variance estimates. This means that in practice we might have to resort to analyzing S_B to get a notion of the Σ_B structure. Fortunately, experience shows that when it is possible to analyze both matrices, similar results are obtained. An alternative to this analysis is to use MFA with an unrestricted Σ_W matrix (see also Longford & Muthen, 1990), only testing the restrictions on Σ_B.

(5) – The next set of steps uses the outcomes of the earlier steps to specify a sequence of MFA's. As is shown in (17), these analyses make use of S_{PW} and S_B simultaneously. The computations are not complicated by a non–positive definite Σ_B estimate since this matrix only appears in the sum $\Sigma_W + c\,\Sigma_B$.

5. Factor analysis results

The analysis steps suggested above will now be applied to the eight achievement variables at both pretest and posttest, followed by multilevel factor analyses for pretest and posttest. Finally we will carry out a longitudinal MFA for both pretest and posttest. The longitudinal analysis gives an efficient way of studying change in between and within variance components over time. A two–level MFA for students within classes will be used in all cases. The school level will be ignored here for simplicity. Since there are only two classes per school and the school variance proportions are relatively small, this clustering effect should not

seriously bias the results.

5.1 Analysis of pre and posttest

Factor analyses of educational data to date have routinely ignored the multilevel character of data which frequently is at hand. Because of this it is of great interest to compare the results of conventional analyses with those of the more refined MFA. This means that we will look in some detail at the results of the analysis steps previously outlined.

Table 3 shows the chi–square tests of model fit and estimated item characteristics for pretest, while Table 4 gives the same values for posttest (standard errors of estimates will not be given in this paper since all models presented show parameters significantly different from zero due to the large sample size). The univariate skewness and kurtosis values do not indicate substantial deviations from the assumed normality, which might have been the case given the small number of items forming the subscores. The step 1 conventional ML analysis of S_T gives a reasonable fit at both pretest and posttest given the large sample size of 3,724. This sample size makes the power of the test very big and rejection at the 5% level may reflect trivial deviations from the model. The step 2 intraclass correlations for the eight items are in the range .18 – .39 at pretest and .24 – .40 at posttest. The values increase over time for all variables and particularly for EQEXP and PFIGURE. This is in line with the random effects analysis of variance results discussed in connection with Table 2. Individual–level measurement error probably deflates the intraclass correlations. The fact that they are still big certainly makes it worthwhile to proceed to step 3.

The third step carries out the analysis of the pooled–within matrix S_{PW}. For both pretest and posttest the conventional ML analysis gives a worse fit for S_T than S_{PW} for the one–factor model. The difference in number of observations is negligible, N = 3,724 versus N – G = 3,527, and cannot alone explain the difference. The worsening of fit is expected given the large size of the intraclass correlations and the large average class size of about 20. Judging from the S_{PW} analysis the within part of the model has a very good fit to the one–factor model given the large sample size. It is also interesting to note from Tab es 3 and 4 that

the conventional analysis of S_T strongly overestimates the reliabilities of the variables relative to the S_{PW} analysis. The S_{PW} analysis adjust for differences in class means. Heterogeneity in the means across classes increases the reliable part of the variation which inflates the reliabilities (see also Muthen, 1989, pp. 559–560). The S_T reliabilities may be correct for an inference to this particular mixture of class means, but is not correct for the student scores in any of the classes. This is further discussed below in connection with the MFA results. It appears that the conventional S_T factor analysis of students sampled within classes can be quite misleading.

In the fourth step we investigate the between structure. The estimated Σ_B was scaled to a correlation matrix and subjected to ordinary exploratory factor analysis by unweighted least squares. Judging from the eigenvalues, a one–factor model holds at both pretest and posttest. For pretest the first four values are 7.08, 0.26, 0.21, 0.17, while for posttest they are 6.79, 0.30, 0.25, 0.21. The two–factor solutions had no interpretable structure. The analysis of the correlation matrix corresponding to S_B gave similar results. The estimated loadings are rather close to those obtained via the estimated Σ_B, although somewhat lower overall.

The first MFA step uses a one–factor model for both within and between since this was suggested in previous steps. This is the model of Figure 1. As in conventional factor analysis the metric of each factor has to be determined and this is done by fixing the between and within loadings for RPP to unity. The chi–square test of model fit is 106.16 and 116.00 for pre and post with 40 degrees of freedom. Given the sample size if 3,724 this is taken as a good fit. The S_{PW} analysis fitted the within part of the model with 20 degrees of freedom, which may be viewed as an analysis with no between structure imposed. The addition of the between structure in the MFA adds about 50 – 60 chi–square points for an additional 20 degrees of freedom. This increase does not seem unduly large for the sample size. It is interesting to note the perfect agreement to two digits in the estimated within reliabilities for MFA and S_{PW}. This is because the MFA estimation of Σ_W is largely determined by the second group in the fitting function of (17) due to the large number of students per class.

Tables 3 and 4

The MFA within reliabilities (see definitions in Section 3.2) are very low as should be expected given the small number of items comprising each subscore. As expected the highest reliability values occur for the arithmetic topics of RPP and FRACT, perhaps not only because these subscores consist of more items than the others, but also because these topics have higher eighth grade OTL (see Table 1). There is a strong increase over time, particularly for EQEXP and PFIGURE. These correspond to new topics at pretest for many eighth graders, whereas they have been better covered at posttest.

Note also that the S_{PW} analysis gives reliabilities which agree with the within values of the MFA to two digits. The higher S_T values observed above may be viewed in terms of the MFA model of (5). For simplicity, assume that to a reasonable approximation λ_{Bj} equals λ_W. Then the reliable part of the S_T variance is modelled as $\lambda^2_j (\sigma^2_{\eta B} + \sigma^2_{\eta W})$ while the error variance sums the between and within errors. The reliable part is an increase compared to WF of (5), which taken together with a relatively small between error variance results in the S_T reliability overestimation. In this application, both the pretest and posttest data led to a rejection of the test of equality of between and within loadings. This may be due to the large sample size, however, since the pattern of estimated loadings is very similar.

The between reliabilities are very high and do not change much from pretest to posttest. It is interesting to note that the largest reliability increase occurs for the algebra content of EQEXP, meaning that EQEXP becomes a better measurement at posttest of the dimension that makes classrooms differ. On the whole, however, the indicators of the between factor are very homogeneous adding very little variation around the general dimension. It may be noted that the step 4 analysis of the estimated Σ_B gave between reliabilities which are almost identical

to the MFA results. Step 4 analysis based on S_B, however, gives consistently lower between reliabilities.

It is interesting to return to the question of variance decomposition discussed in Sections 2 and 3.1 and compare the results of conventional random effects analysis of variance with those of the multilevel factor analysis. Tables 3 and 4 also include the estimated true intraclass correlations. The true intraclass correlation (9) for each variable is calculated as the error–free variance ratio BF/(BF+WF) using the notation of (5). The values are around 0.6 with little difference between the pre and posttest results[1]. This value should be compared to the observed variable intraclass correlations, or proportion between variation, of Table 2 which were in the range 0.2 – 0.4. In this way between class variation becomes relatively more important when purging the measurement error in the scores. This is in line with the expectation that measurement error inflates the within variation. While Table 2 shows a slight increase over time in the proportion between for all variables, the true intraclass correlations of Tables 3 and 4 show a slight decrease for several variables.

5.2 Longitudinal model

Of particular interest in this achievement analysis if the change from pretest to posttest in variance contributions. A more efficient use of the data is to perform a simultaneous analysis of pretest and posttest data. Such a longitudinal model also makes it possible to study change in variance contributions due to the between and within factors, ensuring that these factors are measured in comparable metrics over time. Figure 2 outlines the longitudinal MFA model.

Figure 2

Drawing on the pre– and post–test analyses the longitudinal model specifies one between factor and one within factor for each of the two time points. The two between factors are allowed to be correlated and so are the two within factors. The variables may also be allowed to correlate over time via correlated measurement errors. The need for correlated individual–level errors is often

found in conventional covariance structure analyses where the same instrument is repeatedly administered. Here we extend this to correlation of between errors. For example, if at pretest classroom differences in algebra were beyond what could be explained by the general level of the pretest (a proxy for the between factor) this algebra difference might prevail at posttest leading to a between error correlation.

To be able to study change in factor–related variances over time it is necessary to specify the same metric for both the between and the within factor over time. To this aim we want to restrict the loadings to be equal over time at both the between and within level. We know, however, a priori that some math topics become much more familiar to the students over the course of the eighth grade and therefore may lead to different measurement properties of subscores over time. Coverage of other topics does not change as much over time. This is also supported by the OTL differences and the pre–post differences in estimated reliabilities. For such reasons we will allow the three variables of EQEXP, AREAVOL, and PFIGURE to have different loadings over time, while the other loadings are restricted to be equal, apart from the fixed loading for RPP.

Table 5 shows the different steps of MFA model fitting. Results from analyzing S_T and S_{PW} are given in addition to MFA results for comparison. In the baseline model 1 there is no loading invariance imposed and no error correlations are included. The fit is poor for this model. Adding loading invariance over time also gives a poorly fitting model. Model 3, using partial loading invariance, improves the fit dramatically with a loss of only three degrees of freedom. It is interesting to note the large difference in fit between using S_T and using S_{PW}. Again, using S_{PW} is correct given the multilevel model. This analysis points to a good individual–level fit. Model 5 adds between correlations to the MFA model resulting in a significant improvement in fit. Comparing the model 4 result for S_{PW} with the model 5 result for MFA shows that doubling the number of degrees of freedom by addition of the between structure approximately doubles the chi–square value. We conclude that the MFA model 5 fits reasonably well in both its within and between part given the large sample size.

Table 5

Although all significant, the within error correlations are rather small, in the range 0.07 –0.21. The between error correlations are considerably larger, in the range 0.25 – 0.78. The factor correlations point to a strong linear relationship over time, .80 for within and .93 for between. The S_T and S_{PW} analyses gave factor correlations of 0.88 and 0.80, respectively. The S_T value overestimates the within value while the S_{PW} value is accurate. The percentage factor variance change from the longitudinal MFA model obtain the values 26% and 40%, respectively for between and within. These changes are discussed in more detail in connection with Table 6 below.

Table 6 gives estimated variance ratios for the eight subscores in the form of reliabilities, true intraclass correlations (BF / (BF + WF)), and true increases from pre to post. The reliabilities and the true intraclass correlations are similar to those of Tables 3 and 4 for the separate analyses of pre and post.

Table 6

A new indication of the inflation of within variation due to measurement error is seen in the two right–most columns of Table 6. As in Table 2 these display the increase in variance from pretest to posttest relative to the pretest value. In Table 6 the error–free values BF and WF are used. Comparing with the two right–most columns of Table 2 we find a very different picture. Overall, in Table 6 the between variance increase over time is not as large and the within increase is much larger. In fact, the within increase is the largest with one exception. The Table 2 results are distorted by individual–level measurement errór. Despite an increase over time in true within variance due to the factor the decrease in the measurement error variance over time substantially dampens the total within variance increase. We now find that the error–free within variance increases dramatically over time, or, in other words, within–class student heterogeneity

increases dramatically. This may be due to increasing individual differences due to increasing learning opportunities.

5.3 Unreliability sensitivity analyses

The multilevel factor analysis results provided estimates of the true, or error–free, proportions for each variable of between to total variance using the BF / (BF+WF) ratio of variances due to the factor. It also provides an error–free assessment of change from pre to post in between and within variance due to the factor. The corresponding observed variable quantities from analysis of variance presented in Table 2 were quite different. The Table 2 results pointed to a larger share of within variation and a smaller increase over time in within variation. The differences are hypothesized to be due to measurement error. Using more reliable scores might not make for such a large difference in conclusions. More reliable scores are obtained by the summing of more dichotomous items. Since the assumption of unidimensionality of the items has been supported in the analysis one may contemplate the use of more aggregated subscores. It is of practical interest to get a feeling for how different amounts of aggregation and reliabilities affect the results. Also, the use of different aggregation levels gives a check of the robustness of the MFA results. It is of practical interest to know if the eight variable factor analysis, using variables with very low reliability, gives trustworthy results for the true, error–free estimates.

To study influence of unreliability, RPP and FRACT were combined into a single ARITH(metic) score based on 16 items, EQEXP and INTNUM were combined into an ALG(ebra) score based on 8 items, STESTI and AREAVOL were combined into a MEAS(urement) score based on 7 items, and COORVIS and PFIGURE were combined into a GEOM(etry) score based on 8 items. Also, a total score based on all 39 items will be used.

Table 7 gives analysis of variance estimates for these new scores. Consider first the total score. This score may be viewed as a proxy for the factor in the MFA.

Table 7

The proportion between is .52 at pretest and .53 at posttest. These values are not too far off from the Table 6 average values in the columns BF / (BF + WF) reflecting the reliability of the total score. The percent increase in Table 7 should be compared to the Table 6 column "Error–free % increase". It is clear here that the total score cannot capture the differences in increase for the different parts of the score exhibited in Table 6. In sum, using the total score does not give misleading results, but certainly undifferentiated ones.

The use of the four aggregated scores ARITH, ALG, MEAS, and GEOM turns out to result in biases similar to those that we observed for the eight less aggregated variables. Relative to Table 6, the general picture is still that the proportion between is underestimated, the percent change in between is overestimated, and the percent change in within is underestimated. From a practical point of view it is interesting to note that the biases are quite large even for the 16 item score of ARITH. Going from the 8 item subscores of RPP and FRACT to the 16 items of ARITH decreases the bias considerably but not sufficiently. The effects of unreliability makes it impossible in the analysis of variance to distinguish math topic differences between subscores from differences in the number of items used in the sum to create the score.

Table 8 gives the results of the longitudinal MFA using the new set of four achievement scores. A model analogous to model 5 in Table 5 is used. The four–variable results of Table 8 give a picture similar to the eigth–variable results of Table 6. In this data set it is clear that, unlike anova, MFA is not sensitive to the level of variable aggregation and reliability.

Table 8

6. Conclusions

Multilevel factor analysis has been shown to give new types of useful information on educational test scores. Using the structure of the sample design, the effects of clustered (nested) observations is not only taken into account but also modeled in interesting ways to shed light on within and between class variance components and their changes over time.

From a substantive point of view, it was found that the strong elements of tracking in eighth grade math classes makes for between–class variation in the achievement scores which is about as large as the within–class student variation. At the same time, however, within–class variability increases much more substantially than between–class variation over the course of eighth grade. Increase in both between and within variation is particularly dramatic for algebra topics related to equations and expressions and geometry topics related to plane figures.

From a methodological point of view, several interesting findings emerged. Due to unreliability in the observed scores, the results obtained by analysis of variance are quite different from those of MFA. Anova substantially underestimates the intraclass correlation, or the proportion of between–class variation and substantially underestimates the increase over time in within–class variation. For trustworthy anova results on sums of dichotomously scored items large sets of items are needed and this may preclude differentiation of subtopics. It is also clear that conventional factor analysis of the usual sample covariance matrix gives distorted results. The chi–square test of fit is inflated and estimates are severly biased. MFA is a readily available technique since it can be carried out with standard structural equation software. Given this, it is hoped that educational researchers quickly adopt these exciting new analysis tools. However, MFA is not a small–sample technique. In particular, MFA calls for data that have a sizeable number of groups, preferrably at least about 100. As was pointed out by Cronbach (1976), if cost permits it may be better to observe fewer students per class in favor of including more classes.

Several extensions of the MFA models studied here are available. In addition to class–level components of student–level variables one may include class–level variables. For example the class–level OTL information can be incorporated to

explain the class–level student achievement variation (see Muthen, 1990). The modeling is not limited to factor analysis, but structural equation models can also be analyzed (see Muthen, 1989). More than two levels of nesting can be incorporated. All of these extensions fit into conventional software using the MUML–type estimator. The MFA techniques are of course not limited to educational data and students observed within classes and schools but can be used in any situation where cluster sampling has been employed, for example with geographically determined groups of households.

Footnotes

1. The metric of the factors is noteworthy. The choice of variable for the loading fixed at unity does influence the ratio of between factor variance divided by the sum of between and within factor variance. While in Tables 3 and 4 the use of RPP gives the factor ratio value of RPP, the use of STESTI would instead give STESTI's ratio. This is the case whenever the set of loadings differ on the between and within level. On the other hand, the corresponding ratios for the variables, given in these tables, are invariant in this regard as are all other values given in the tables.

References

Bock, R. D. (1989). Multilevel analysis of educational data. San Diego, CA: Academic Press.

Bohrnstedt, George W. (1983). Measurement. In P. H. Rossi, J. D. Wright, & A. B. Anderson, (Eds.) Handbook of survey research. New York: Academic Press.

Bollen, Kenneth A. (1989). Structural Equations with Latent Variables. New York: Wiley.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. Review of Research in Education, 8, 158-233.

Burstein, L. (Ed.) (1990). The IEA Study of Mathematics III: Student Growth and Classroom Process. London: Pergamon Press.

Cronbach, L. J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. Unpublished manuscript, Stanford University, Stanford Evaluation Consortium, School of Education.

Crosswhite, F.J., Dossey, J. A., Swafford, J. O., McKnight, C. C., & Cooney, T. J. (1985). Second international mathematics study: Summary report for the United States. Champaign, IL: Stipes.

Fisher, R.A. (1958). Statistical Methods for Research Workers (13th Ed.). New York: Hafner Publishing Co. Inc.

Goldstein, H. I., & McDonald, R. P., (1988). A general model for the analysis of multilevel data. Psychometrika, 53, 455-467.

Haggard, E.A. (1958). Intraclass Correlation and the Analysis of Variance. New York:. Dryden Press, Inc.

Harnquist, K. (1978). Primary mental abilities of collective and individual levels. Journal of Educational Psychology, 70, 706-716.

Joreskog, K. G. & Sorbom, D. (1979). Advances in Factor Analysis and Structural Equation Models. Cambridge, MA: Abt Books.

Koch, G.G (1983). Intraclass correlation coefficient. Encyclopedia of Statistical Sciences, 4, 212-217.

Longford, N.T., Muthén, B. (1990). Factor Analysis for Clustered Observations. The University of California, Los Angeles.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. British Journal of Mathematical and Statistical Psychology, 42, 215-232.

Muthén, B. (1988) Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study. Graduate School of Education, University of California, Los Angeles.

Muthén, B. (1989). Latent variable modeling in heterogenous populations. Presidential address to the Psychometric Society, July, 1989. Psychometrika, 54, 557-585.

Muthén, B. (1990). Mean and Covariance Structure Analysis of Hierarchical Data. Paper presented at the Psychometic Society meeting in Princeton, N.J., June 1990. UCLA Statistical Series #62.

Muthén, B. & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. Invited paper for the conference "Multilevel Analysis of Educational Data," Princeton NJ, April 1987. To appear in an edited book with the same name (Ed. D. Bock).

Rakow, E.A., Airasian, P.W. & Madaus, G.F.. (1978). Assessing School and Program Effectiveness: Estimating Teacher Level Effects. Journal of Educational Measurement, 15(1)., 15-21.

Raudenbush, S., & Bryk, A. (1988). Methodological advances in studying effects of schools and classrooms on student learning. Review of Research in Education, 1988.

Schmidt, W.H. (1969). Covariance structure analysis of the multivariate random effects model, unpublished doctoral dissertation. University of Chicago.

Schmidt, W. & Wisenbaker, J. (1986). Hierarchical data analysis: an approach based on structural equations. CEPSE, No.4., Research Series, Department of Counseling Educational Psychology and Special Education.

Schmidt, W., Wolfe, R.G. & Kifer, E. (1990). The Identification and Description of Student Growth in Mathematics Achievement. In Burstein, L. (Ed.), The IEA Study of Mathematics III: Studes

Skinner, C.J., Holt, D., Smith, T.M.F. (1989). Analysis of Complex Surveys. Chichester: John Wiley & Sons.

Wiley, D.E. & Bock, R.D. (1967). Quasi-experimentation in Educational Settings: Comment. The School Revew, 75(4), 353-366.

Winer, B.J., (1971). Statistical Principles in Experimental Design. New York: McGraw-Hill.

Table 1

Performance and opportunity to learn for eight math achievement subscores

Subscore	Number of items	Pretest		Posttest		Opportunity to learn	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
RPP	8	3.41	2.12	4.19	2.30	7.48	1.19
FRACT	8	3.28	1.96	4.09	2.14	7.52	0.83
EQEXP	6	2.38	1.42	2.98	1.63	3.96	1.82
INTNUM	2	0.65	0.70	1.04	0.79	1.88	0.41
STESTI	5	2.91	1.30	3.14	1.37	4.09	1.18
AREAVOL	2	0.64	0.74	0.91	0.81	1.74	0.54
COORVIS	3	1.14	0.91	1.51	0.98	1.61	0.95
PFIGURE	5	1.63	1.27	2.32	1.47	2.91	1.45

Table 2

Variance decomposition of achievement scores
(percentages of total variance in parenthesis)

	Number of items	Pretest				Posttest				% Increase	
		School	Class	Student	Prop between	School	Class	Student	Prop between	Between	Within
RPP	8	.189* (4.2)	1.353 (29.9)	2.990 (66.0)	.34	.638 (11.8)	1.446 (26.7)	3.326 (61.5)	.38	35.1	10.9
FRACT	8	.337* (8.8)	1.123 (29.4)	2.366 (61.8)	.38	.557 (11.9)	1.349 (28.9)	2.767 (59.2)	.41	30.5	16.9
EQEXP	6	.089* (4.4)	.454 (22.5)	1.473 (73.1)	.27	.260 (9.7)	.781 (29.1)	1.646 (61.3)	.39	91.7	17.7
INTNUM	2	.020* (4.0)	.107 (21.2)	.358 (70.9)	.29	.053 (8.3)	.142 (22.3)	.442 (69.4)	.31	53.5	23.5
STEST	5	.159 (9.1)	.421 (24.2)	1.163 (66.7)	.33	.179 (9.3)	.485 (25.2)	1.258 (65.5)	.34	14.5	8.2
AREAVOL	2	.017* (3.1)	.077 (14.1)	.451 (82.8)	.17	.062 (9.6)	.094 (14.6)	.490 (75.9)	.24	66.0	8.6
COORVIS	3	.028* (3.4)	.145 (17.5)	.656 (79.1)	.21	.073 (7.6)	.202 (21.2)	.680 (68.3)	.32	59.0	3.7
PFIGURE	5	.062* (3.9)	.301 (19.0)	1.224 (77.1)	.23	.274 (12.7)	.437 (30.1)	1.451 (67.1)	.33	95.9	18.5

* Not significant at 5% level

Table 3
Pretest analysis results

Model tests

Method	Chi-square	D.F.
S_T	83.71	20
S_{PW}	58.29	20
MFA		
MUML	106.16	40
FIML		40

Item Characteristics

				Reliability				
						MFA		MFA
	Skewness	Kurtosis	Intraclass correlation	S_T	S_{PW}	Within	Between	BF/(BF+WF)
RPP	.38	-.68	.34	.61	.44	.44	.96	.52
FRACT	.37	-.57	.39	.60	.38	.38	.97	.61
EQFXP	.23	-.57	.27	.36	.18	.18	.83	.64
[illegible]	.60	-.80	.27	.34	.18	.18	.81	.63
STESTI	-.24	-.64	.32	.44	.25	.25	.86	.61
AREAVOL	.68	-.89	.18	.29	.18	.18	.82	.50
COORVIS	.38	-.70	.21	.34	.18	.18	.92	.57
PFIGURE	.61	-.21	.24	.32	.17	.17	.78	.59

Table 4
Posttest analysis results

Model tests

Method	Chi-square	D.F.
S_T	88.59	20
S_{PW}	57.45	20
MFA		
MUML	116.00	40
FIML		40

Item Characteristics

	Skewness	Kurtosis	Intraclass correlation	Reliability				
				S_T	S_{PW}	MFA Within	MFA Between	BF/(BF+WF)
RPP	0.03	-1.07	.38	.68	.52	.52	.97	.53
FRACT	-0.01	-0.92	.40	.68	.49	.49	.98	.57
EQEXP	-0.02	-0.89	.38	.55	.32	.32	.92	.64
INTNUM	-0.07	-1.41	.30	.43	.25	.25	.88	.61
STESTI	-0.44	-0.62	.33	.52	.34	.34	.89	.56
AREAVOL	0.16	-1.44	.25	.38	.23	.23	.84	.54
COORVIS	-0.03	-1.00	.30	.42	.26	.26	.80	.55
PFIGURE	0.15	-0.94	.33	.46	.31	.31	.77	.54

Table 5

Longitudinal model tests

Model	Chi-square (d.f.)		
	S_T	S_{PW}	MFA
1. Baseline	1,041.38 (108)	687.43 (103)	1,101.90 (206)
2. 1 + loading invariance	1,160.47 (110)	734.26 (110)	1,183.59 (220)
3. 1 + partial loading invariance	1,063.31 (107)	696.31 (107)	1,117.16 (214)
4. 3 + correlated within errors	369.64 (99)	221.67 (99)	589.59 (206)
5. 4+ correlated between errors	-	-	450.63 (198)

Table 6
Item characteristics estimated from the
longitudinal multilevel factor analysis model

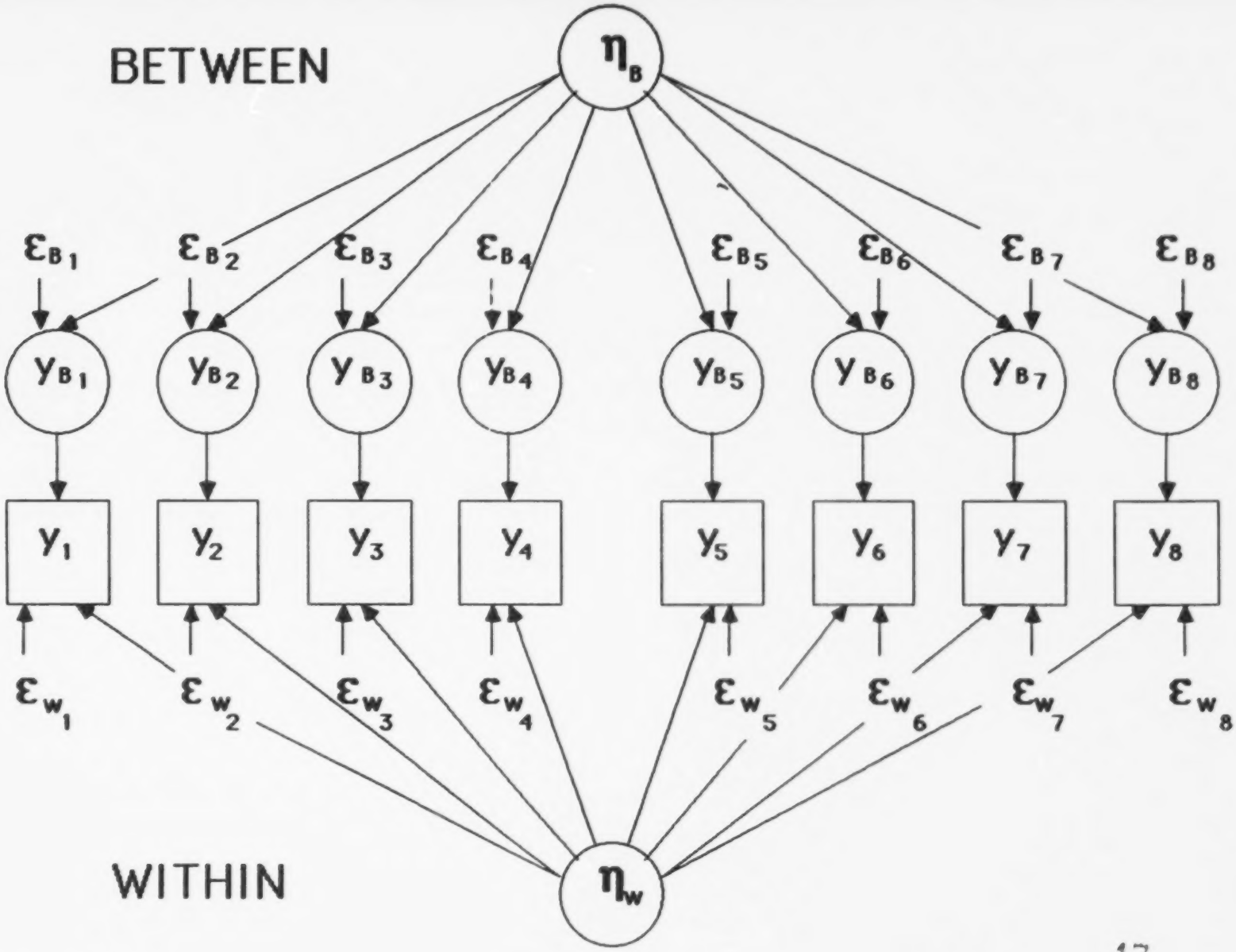
	Reliability				BF/(BF+WF)		Error-free % increase	
	Pre		Post		Pre	Post		
	Between	Within	Between	Within			Between	Within
RPP	.97	.43	.96	.53	.54	.51	25.7	40.0
FRACT	.96	.40	.97	.48	.60	.57	25.7	40.0
BQEXP	.82	.17	.93	.32	.65	.64	113.0	117.0
INTNUM	.82	.19	.87	.23	.63	.61	25.7	40.0
STESTI	.85	.26	.89	.34	.58	.56	25.7	40.0
AREAVOL	.80	.16	.84	.24	.51	.54	76.0	55.4
COORVIS	.91	.19	.80	.26	.57	.55	25.7	40.0
PFIGURE	.77	.16	.77	.32	.60	.55	88.7	135.0

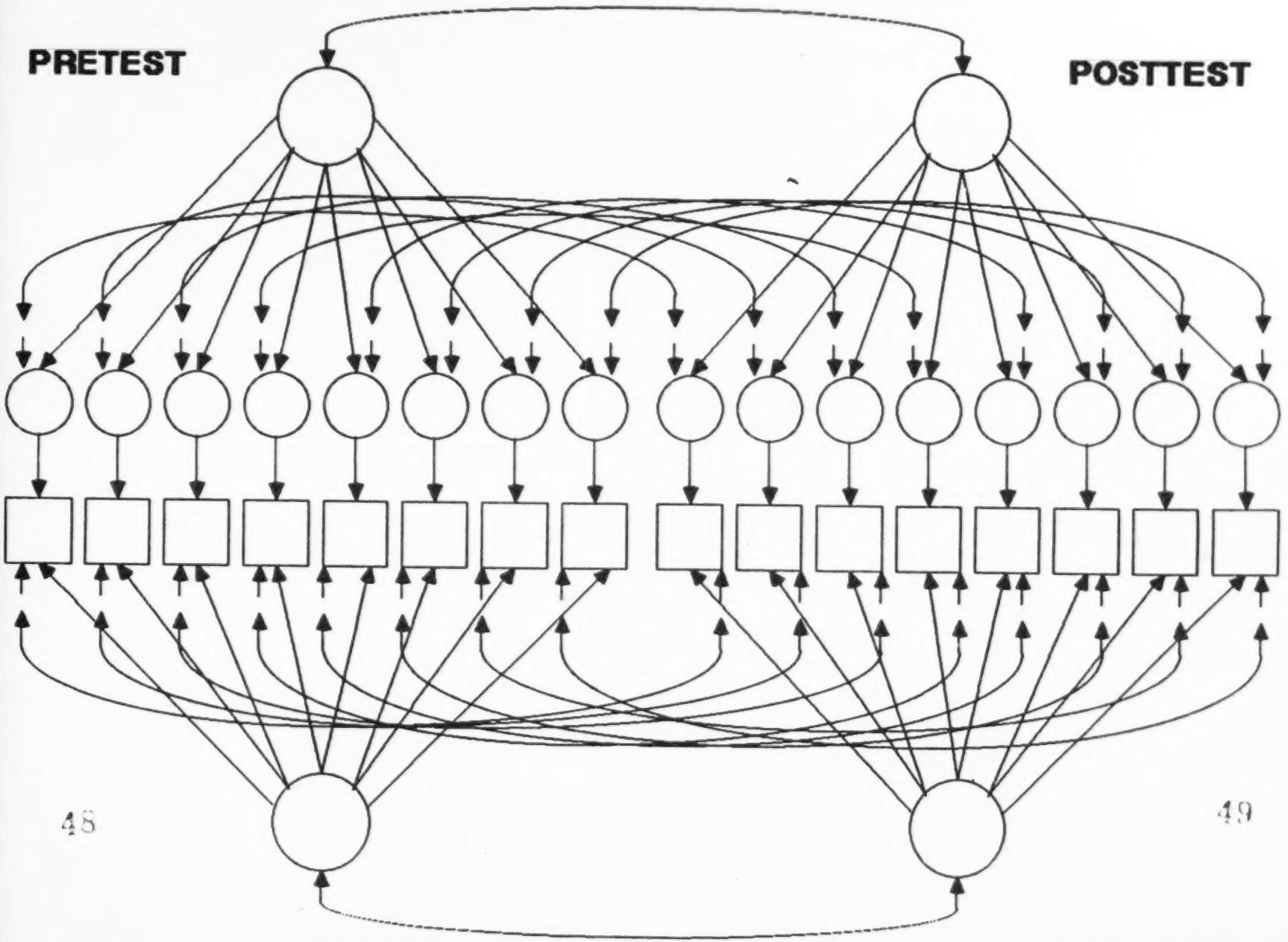
Table 7

Variance decomposition of aggregated achievement scores
(percentages of total variance in parenthesis)

Score	Number of items	Pretest				Posttest				% Increase	
		School	Class	Student	Prop between	School	Class	Student	Prop between	Between	Within
TOTAL	39	4.404* (7.8)	24.377 (43.3)	27.576 (48.9)	.52	12.532 (15.2)	31.039 (37.7)	38.750 (47.1)	.53	51.4	40.5
ARITH	16	1.016* (7.5)	4.919 (36.3)	7.607 (56.2)	.44	2.349 (13.7)	5.589 (32.7)	9.173 (53.6)	.46	33.7	20.6
ALG	8	.172* (5.3)	.964 (29.3)	2.122 (65.1)	.35	.502 (10.9)	1.571 (34.0)	2.543 (55.1)	.45	82.5	19.8
MEAS	7	.229* (7.8)	.815 (27.9)	1.882 (64.3)	.36	.430 (12.2)	.949 (26.9)	2.155 (61.0)	.39	32.1	14.5
GEOM	8	.131* (4.1)	.842 (26.1)	2.247 (69.8)	.30	.585 (13.3)	1.129 (25.6)	2.701 (61.2)	.39	76.2	20.2

* Not significant at 5% level


BETWEEN
η_B
ε_{B_1}
ε_{B_2}
ε_{B_3}
ε_{B_4}
ε_{B_5}
ε_{B_6}
ε_{B_7}
ε_{B_8}
y_{B_1}
y_{B_2}
y_{B_3}
y_{B_4}
y_{B_5}
y_{B_6}
y_{B_7}
y_{B_8}
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
ε_{w_1}
ε_{w_2}
ε_{w_3}
ε_{w_4}
ε_{w_5}
ε_{w_6}
ε_{w_7}
ε_{w_8}
WITHIN
η_w


PRETEST
POSTTEST